CS410 Project Proposal

Team BXB

1. ***What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.***

Xiaofan Liu / xl109@illinois.edu / Captain
Boyu Pang / bpang4@illinois.edu
Brittany West / bnwest2@illinois.edu

2. ***What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?***

Our topic is “Campuswire and Coursera Query Link”, which falls under the “Intelligent Browsing” theme. The goal is to develop a program to present users with relevant information from Coursera related to questions that are posted on Campuswire. This program would facilitate the learning process by providing relevant ranked documents from Coursera to the user, which reduces the quantity of repeated questions on Campuswire that could quickly crowd out other questions/answers.

The first problem is that due to limited searching functionality, sometimes it is difficult to find answers quickly on Coursera. The second is that there are many questions posted on Campuswire which are answered in the lecture transcripts, syllabus or reading materials, but there is no easy way to identify the exact place to go in Coursera for the answer.

3. ***Briefly describe any datasets, algorithms or techniques you plan to use***

Dataset: CS410 course data on Coursera (or a downloaded dataset that contains all information for this course)

Techniques:

- We will browse Campuswire and compile a list of common questions/topics/themes.
- Use those themes to form the basis of our query set (i.e use queries derived from our list to search the provided Coursera data).
- Preprocess the Coursera data by separating it into distinct documents.

Algorithms:

- For our source code:
    - We will explore incorporating PLSA to help improve our ranking algorithm. In our PLSA algorithm we can implement the latent dirichlet allocation.
    - For our ranking function, we can evaluate multiple algorithms to determine which fits our project the best.

- We can evaluate our results/ranking function by manually identifying relevant documents on a set of queries, and then applying MAP.
- Our output will include relevant documents with the first matching keywords of each document.

4. ***How will you demonstrate that your approach will work as expected? Which programming language do you plan to use?***

We will demonstrate by walking through our code and doing a test run with mock-up questions/queries for the course. We will use Python.

5. ***Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.***

Our workload and tasks are listed as follows:

- Brainstorming and team meetings (5 hours)
- Scraping and Data cleaning, including structuring and cleaning the coursera documents, gathering queries to test our ranking function (10 hours)
- Researching, implementing, fine tuning ranking functions (20 hours)
- Testing ranking functions and improving them (10 hours)
- Documenting code and writing project updates (5 hours)
- Preparing and creating presentation (10 hours)

Total hours : 60